DOCUMENT RESUME

ED 185 160 UD 020 354

AUTHOR Baker, Eva
TITLE What Do We Know About Teaching and Learning in Urban Schools? Volume 4: Achievement Tests in Urban Schools: New Numbers.
INSTITUTION Central Midwestern Regional Educational Lab., Inc., St. Louis, Mo.
SPONS AGENCY National Inst. of Education (DHEW), Washington, D.C.,
PUB DATE 79
NOTE 67p.: Paper prepared for the Urban Education Program, CEMREL, Inc.'s National Conference on Urban Education (St. Louis, MO, July 10-14, 1978). For related documents see UD 020 351-361 and UD 020 363.

EDRS PRICE MF01/PC03 Plus Postage.
DESCRIPTORS *Achievement Tests; Elementary Secondary Education; *Standardized Tests; *State of the Art Reviews; Test Bias; Test Construction; *Testing Problems; Test Reviews; Test Selection; *Urban Schools

ABSTRACT

This paper reviews the research in the area of achievement tests. The findings indicate: (1) there is little evidence that educational decision makers use the large amount of test data collected; (2) a larger number of tests are being required of students for personal certification; (3) reasons for the lack of application of test results include lack of experience and training in test and test data uses, lack of clarity regarding alternative actions implied by test results, and confusion about utility and quality of data; (4) tests are used in ways that directly affect students, assist in school operations, and provide feedback to administrators; (5) controversy exists over the appropriate test type to use for various purposes; (6) estimates of test quality have not been high; (7) emphasis in test development is toward providing description rather than comparative information for student achievement; (8) research in test bias has not been integrated with newer views of test development; (9) teachers in low socioeconomic areas are less satisfied with achievement testing procedures than other teachers; and (10) the cost of testing is too high. Recommendations are made for research and educational personnel. Educators are encouraged to employ consulting firms to develop tailored tests, to educate school personnel and community members on testing and interpretation, and to attempt to limit the number of separate testing activities. (Author/MK)

****
* Reproductions supplied by EDRS are the best that can be made *
* from the original document. *
****

FEB 2 9 1980

WHAT DO WE KNOW ABOUT TEACHING AND LEARNING IN URBAN SCHOOLS?

ED185150

Volume 4

Achievement Tests in Urban Schools: New Numbers

Eva Baker
University of California, Los Angeles

U S DEPARTMENT OF HEALTH, EDUCATION & WELFARE
NATIONAL INSTITUTE OF EDUCATION

THIS DOCUMENT HAS BEEN REPRODUCED EXACTLY AS RECEIVED FROM THE PERSON OR ORGANIZATION ORIGINATING IT. POINTS OF VIEW OR OPINIONS STATED DO NOT NECESSARILY REPRESENT OFFICIAL NATIONAL INSTITUTE OF EDUCATION POSITION OR POLICY

URBAN PROGRAM

CE020354

WHAT DO WE KNOW ABOUT TEACHING AND LEARNING IN URBAN SCHOOLS?

Volume 4

Achievement Tests in Urban Schools: New Numbers

Eva Baker
University of California, Los Angeles

Paper prepared for the National Conference on Urban Education, held in St. Louis, Missouri, July 10-14, 1978.
Conducted by the Urban Education Program, CEMREL, Inc., with support from the National Institute of Education.

CEMREL, Inc., St. Louis, Missouri

FEB 2 6 1980

Prepared by CEMREL, Inc., a private nonprofit corporation supported in part as an educational laboratory by funds from the National Institute of Education, Department of Health, Education and Welfare. The opinions expressed in this publication do not necessarily reflect the position or policy of the National Institute of Education, and no official endorsement should be inferred.

Copyright on this document is claimed only during the period of development, test, and evaluation, unless additional authorization is granted by the National Institute of Education to claim copyright on the final version. For information on the status of the copyright claim, contact either the copyright proprietor or the National Institute of Education.

Copyright 1979, CEMREL, Inc.

Library of Congress catalog card number 79-54733
Manufactured in the United States of America

CEMREL, Inc.
3120 59th Street
St. Louis, MO 63139

Table of Contents

ACKNOWLEDGMENTS..... vii

FOREWORD..... xiii

ACHIEVEMENT TESTS IN URBAN SCHOOLS: NEW NUMBERS..... 1

I. INTRODUCTION..... 1

II. EXISTING EVIDENCE..... 4

III. SYNTHESIS OF EVIDENCE..... 33

IV. RECOMMENDATION FOR RESEARCH PROGRAMS..... 34

V. RECOMMENDATION FOR POLICY AND PRACTITIONERS..... 39

REFERENCE NOTES..... 41

REFERENCES..... 43

Acknowledgments

The Urban Education Program, CEMREL, Inc., would like to thank the National Institute of Education and the Regional Joint Planning Group of the Urban Education Program for its support and assistance in planning the conference.

Conference Chairperson

Harriet Doss Willis
Director
Urban Education Program

Conference Participants

Nettie S. Armmer
Research Associate
St. Louis University

Eva Baker
Director
Center for the Study of Evaluation
UCLA Graduate School of Education

Carrie E. Bash
Division of Community Programs
Urban League of St. Louis

Isabel Beck
Co-director
Language & Communications Unit
Learning Research and Development Center
University of Pittsburgh

Janet Cheatham Bell
Urban/Ethnic Studies Coordinator
Indiana Department of Public Instruction

David Bennett
Deputy Superintendent
Milwaukee Public Schools

Edward Binkley
Director of Research & Evaluation
Metro Schools
Nashville, Tennessee

Milton Bins
Senior Associate
Policy Development
The Council of Great City Schools
Washington, D.C.

Dean Bowles
Director
Home-School-Community Project
Wisconsin Research and Development Center

Ronald A. Boyd
Associate Superintendent
Department of Education
State of Indiana

Emerson Brown
Board of Directors
CEMREL, Inc.

L.D. Buford
Principal
St. Louis Public Schools

William D. Callian, Jr.
Assistant Superintendent
Curriculum Instruction
Memphis City Schools

Angeline P. Caruso
Associate Superintendent
Department of Curriculum &
Instructional Services
Board of Education
City of Chicago

Courtney Cazden
School of Education
Harvard University

Ned Chalker
Teaching & Learning Program
National Institute of Education

Francis S. Chase
Director of Urban Education
Dallas, Texas

Rosemary Christensen
Director of Indian Education
Minneapolis Public Schools

Doris J. Clemons
Region Five Office
Detroit Board of Education
Detroit, Michigan

Frances Coe
Board of Directors
CEMREL, Inc.

David Colton
Graduate Institute of Education
Washington University
St. Louis, Missouri

Lloyd M. Cooke
Vice Chairman
Economic Development Council of
New York City

Dallas Daniels
Director
Equal Employment Opportunity
Indianapolis Department of Public
Instruction

A. W. Dirks
Director
Research, Planning & Development
Wichita Public Schools

Lorenza Dixon
Curriculum Coordinator
Indianapolis Public Schools

John W. Dobbs
Assistant Superintendent
School & Community Affairs
Department of Education
State of Michigan

Domingo Dominguez
Director
Bilingual & Migrant Program
Southwest Educational Development
Laboratory

Ronald Edmonds
Director
Center for Urban Studies
Graduate School of Education
Harvard University

Nancy Evers
Department of Educational
Administration
University of Cincinnati

Joseph L. Felix
Director
Program Evaluation
Cincinnati Public Schools

Marian Kilbane-Flash
Supervisor
Research & Development
Cleveland Public Schools

Anita Ford
Co-chairman
Citizens Education Task Force
St. Louis, Missouri

Ruges R. Freeman
Coordinator
Secondary Student Training
Southern Illinois University
Edwardsville, Illinois

Gary Gappert
New Jersey Department of Education

Booker Gardner
Wisconsin Research & Development
Center for Individualized Schooling
Madison, Wisconsin

Howard Gholson
Policy Analysis
Title I
U.S. Office of Education

Henry Givens
Assistant Commissioner
Department of Elementary &
Secondary Education
State of Missouri

Edmund Gordon
Director
Institute for Urban &
Minorities Education
Teachers College
Columbia University

Milton Goldberg
Educational Organizations &
Institutions
National Institute of Education

John H. Grate
Director
Planning & Development
Cincinnati Public Schools

Eldon Grossner
Manager
LEA Services
Illinois Office of Education

Rex Hagans
Northwest Regional Educational
Laboratory

Robert J. Havighurst
Department of Education
The University of Chicago

Garland M. Hawkins
Director
Music Education
Curriculum Center School Dist. 189
East St. Louis, Illinois

Paul Hill
The Rand Corporation

Keith Hyde
Planning & Development
Des Moines Public Schools

John Ingram
Wisconsin Research & Development
Center for Individualized Schooling

Barbara Jackson
Associate Dean
Department of Administration &
Supervision
School of Education
Atlanta University

James N. Jacobs
Superintendent
Cincinnati Public Schools

Sylvia K. Jones
State & Federal Grants
Toledo Public Schools

Preston C. Kronkosky
Field Services & Dissemination
Southwest Educational Development
Laboratory

Daniel U. Levine
Director
Center for the Study of
Metropolitan Problems
in Education
University of Missouri-Kansas City

Joseph D. Liggins
Assistant Superintendent for Staff Development
Houston Independent School District

Reece Little
Director of Curriculum Development
Jefferson County Public Schools
Louisville, Kentucky

Irene Lober
Superintendent
University City Public Schools
St. Louis, Missouri

William Lobosco
ESEA Title I
U.S. Office of Education

Catherine Lyon
Center for the Study of Evaluation
UCLA Graduate School of Education

Mary Meehan
District Coordinator
The School District of Kansas City

LaMar Miller
Director of Secondary Education
New York University

Len Nachman
Division of Planning & Development
Minnesota Department of Education

Tom Odneal
Director
Title IV
Missouri Department of Education

Joan Orender
Project Coordinator for Distribution of Information
Nebraska Department of Education

George A. Parry
George A. Parry & Associates
River Forest, Illinois

William Phillips
Minneapolis Public Schools

Andrew C. Porter
Institute for Research on Teaching
College of Education
Michigan State University

Electra Price
Oakland Unified School District
Oakland, California

Walter E. Richardson
Supervisor
Compensatory Instruction
Columbus Public Schools

J. Robert Sampson
Assistant Manager
Program Planning & Development
Illinois Office of Education

Judah L. Schwartz
Professor of Engineering, Science and Education
Massachusetts Institute of Technology

Cleona Shortridge
St. Louis Public Schools

Barbara Shull
Conference on Education
St. Louis, Missouri

Edward E. Smith
Right to Read Effort
Washington, D.C.

Douglas Smith
Administrative Assistant to the Governor
State of Michigan

Frank Sobol
Basic Skills Group
National Institute of Education

George Spicely
U.S. Office of Education

Jane Stallings
Stanford Research Institute

Tom Stefonek
Director of Planning & Evaluation Research
Wisconsin Department of Public Instruction

Vera John-Steiner
Department of Educational Foundations
College of Education
The University of New Mexico

Ronald Stodghill
Associate Superintendent
St. Louis Public Schools

Marilyn Suydam
ERIC Information Analysis Center for Science, Math & Environmental Education
The Ohio State University

Maud Thomas
U.S. Office of Education
Kansas City Regional Office

Concepcion Valadez
Graduate School of Education
University of California
Los Angeles

Mamie Wandick
Program Officer
U.S. Office of Education
Kansas City Regional Office

Phyllis Weaver
School of Education
Harvard University

Billy-Belle Weber
Weber, Fuhrmann & Associates
East St. Louis, Illinois

Joyce Weddington
Director
Research and Evaluation
Memphis City Schools

Rachel T. Weddington
University Dean for Teacher Education
The City University of New York

Lores Wells
Harris Teachers College
St. Louis, Missouri

Robert Wentz
Superintendent
St. Louis Public Schools

Barbara Woods
Afro-American Studies
St. Louis University

Saul Yanofsky
National Institute of Education

Savannah M. Young
St. Louis Public Schools

Allen F. Zondlak
Director of Planning
Detroit Public Schools

URBAN EDUCATION PROGRAM STAFF

John Bass
Resource Coordinator

Alfreda Brown
Resource Specialist

Pamelia Coe
Research Associate

Eileen Hamilton
Secretary

Nellie H. Harrison
Dissemination-Curriculum Specialist

Donald Miller
Research Advisor

Harriet Doss Willis
Director

Astacia Wright
Research Assistant

STAFF FOR WHAT DO WE KNOW ABOUT TEACHING AND LEARNING IN URBAN SCHOOLS?

Harriet Doss Willis
Editor

Margaret Solomon
Manuscript Editor
Production Coordinator

Linda Cambell
Office Manager

Marlene Dandridge
Floor Clerk

Arneita Gray
Secretary

Angela Reed
Secretary

Dian Taylor
Secretary
Office Manager

Karen Temmen
Secretary

Foreword

This monograph is one of a series of papers on the issues of instruction and learning in urban schools that were presented at a conference held July 10-14, 1978, in St. Louis, Missouri, sponsored by the Urban Education Program, CEMREL, Inc., and supported by the National Institute of Education. It is our hope that both the conference and the series of fourteen monographs that resulted from it will assist educational researchers and school practitioners in identifying and analyzing instructional and learning problems in urban school settings, and will contribute to the development of strategies for the improvement of schooling for students.

Titled "What Do We Know About Teaching and Learning in Urban Schools?", the 1978 conference focused on an examination of research findings on the teaching-learning process in areas most germaine to and under control of the schools. The key problems to which the papers are addressed are the assessment of learning outcomes and the analysis of the relationships between instructional and other inputs and learning outcomes. Important contextual issues not under the full control of schools, such as community and parental involvement and the impact of federal and state governments in improving educational opportunities, were addressed at the conference as well.

The conference was conceived, in part, as an initial step in the development of a state-of-the-art review of the most critical issues faced by educators in urban schools who struggle with the morass of general educational problems, exacerbated by decreasing public confidence in large city schools, declining student enrollment, and increasing numbers of poor and minority students. While the conference does not exhaust discussion of the problems

that are particularly significant in the urban school setting, it forms a convenient focus for viewing the larger issues affecting education in our cities.

The educational innovation and change growing out of more than ten years of effort since the passage of the Elementary and Secondary Education Act of 1965 -- the primary stimulus for federal support of research and development in education -- has had the least impact on the student populations typical of the urban school setting. With this concern as a primary consideration, ten months prior to convening the conference, CEMREL's Urban Education Program had begun work with a network of twenty big-city school districts and state departments of education in Illinois, Indiana, Iowa, Kansas, Kentucky, Michigan, Minnesota, Missouri, Nebraska, Ohio, Tennessee, and Wisconsin. Thus, the conference and the research findings of national authorities was undertaken to lay a foundation for future work in urban educational research and development and to undergird present efforts at improvement of schools in the region.

The Conference Design

To guide the development of the presenters' preparation of their papers, topical questions were prepared for them by Urban Education Program Staff and a paper format was recommended. Each presenter prepared a full paper based on the topical questions and results were summarized by the writers during general sessions of the conference. Because of a wide range of perspectives held on the conference issues, discussions precipitated by reactor panels composed of other prominent researchers, practitioners, and community leaders in education followed each general conference session. More than one hundred educational researchers and practitioners from across the country attended the conference. The invited participants came from school districts and state departments of education from nineteen states; thirteen institutions of higher education; four federal agencies; numerous educational laboratories and research and development centers; and seven social service agencies.

The conference was divided into two types of sessions: general sessions, wherein the nationally recognized speakers who wrote papers presented their summaries; and forum sessions, wherein reactors helped analyze the presentations and examine the concrete recommendations put forth by the general-session speakers. All participants were encouraged to present questions and reactions during these forum sessions.

Francis S. Chase, Director of Urban Education Studies, Dallas, Texas, gave the keynote address at the opening session of the conference, a banquet on Monday evening, July 10. His topic, "Promising Developments in Urban Education," summarized field studies that he was conducting under sponsorship of the Council of Great Cities Schools and the University Council for Educational Administration under a grant from the Spencer Foundation. Chase is Professor Emeritus of the University of Chicago, where he also served as Chairman and Dean of Education for ten years.

The Conference Topics

Each of the presenters had been asked to design a paper by responding to salient topical questions, drawn up by CEMREL's Urban Education Staff. The questions that were posed originally for each of the writers are presented below. Most of the resulting papers were developed into monographs for this series.

Compensatory Education: A Congressional Perspective Christopher T. Cross Minority Staff Director Committee on Education and Labor U.S. House of Representatives.	What is the current state of the evidence concerning projected changes in legislative mandates, and what may we expect related to changes in policy governing federally funded compensatory education programs?

Effective Teacher-Support Systems to Improve Instructional Conditions in Schools Joseph D. Liggins Assistant Superintendent for Staff Development Houston Independent School District	What teacher support systems are most effective in improving instructional conditions for students? What is the role of in-service training, and what types of in-service training can be recommended to urban educators? What other resources--such as additional curricular materials, support from teacher aides and volunteers, time re-allocations, team teaching arrangements, etc.--can be demonstrated to be effective in improving students' learning opportunities? What is the state of the evidence?
Assessment That Respects Complexity in Individuals and Programs Judah L. Schwartz Professor of Engineering, Science and Education Massachusetts Institute of Technology	How can we best measure and analyze the effects of instruction on students' learning in the major school subjects: What are the most appropriate times and strategies for measuring the effects of schooling on student learning? What do we learn from means of assessment other than standardized achievement testing? When and under what circumstances is it most appropriate to use alternative or complementary strategies for assessing program effectiveness?
Achievement Tests in Urban Schools: New Numbers Eva Baker Director Center for the Study of Evaluation University of California, Los Angeles	How can we best measure and analyze the effects of instruction on students' learning in the major school subjects? What are the most appropriate times to measure the effects of schooling on student learning? What are the cautions which should be taken into account before any attempt to evaluate the effectiveness of programs primarily on the basis on standardized achievement data, especially when the number of individual implementations is large and the actual conditions under which the implementations took place are not well documented?
Summer Drop-off and The Effectiveness of Compensatory Instruction Paul T. Hill The Rand Corporation	What has the era of compensatory education program development contributed to knowledge about achievement in the basic school subjects in urban schools? Where is the evidence? What, if anything, does the evidence indicate about methods for improving basic skills achievement? What, if any, recommendations can be made to school district planners on the basis of existing evidence?

Instructional Ingredients for the Development of Beginning Reading Competence Isabel Beck Co-Director Learning Research and Development Center University of Pittsburgh	Much research, program development, and implementation of efforts to improve achievement in basic school subjects has been concentrated in the area of reading. What approaches to instruction in reading have been successful in increasing achievement? Is there evidence that certain approaches are best suited to the needs of particular populations? Do these effects generalize to other areas of school learning, such as mathematics, social studies, and science?
A Discussion of The Literature and Issues Related to Effective Schooling Ronald Edmonds Director Center for Urban Studies Graduate School of Education Harvard University	Are there schools that have demonstrated a higher degree of effectiveness in the delivery of instruction in the basic school subjects than others serving similar student populations (in terms of socioeconomic backgrounds, family characteristics, minority populations, etc.)? What characteristics distinguish these schools from less successful schools in similar situations? What is the evidence that the explanatory characteristics have been correctly identified, and how strong is that evidence?
The Educational Promise of Cultural Pluralism Vera John-Steiner and Larry Smith Department of Educational Foundations College of Education The University of New Mexico	How can we approach problems of teaching and learning in urban schools so that the racial and cultural diversity of the school population is recognized, respected, and utilized constructively, to the extent possible? How do we differentiate between basic academic skills, which all students need to acquire for successful participation in our hetergeneous society, and curriculum decisions which can and should be tailored to the needs and interests of different sub-cultural groups? Do students benefit from participation in racially and culturally hetero-geneous classrooms, or perhaps from certain kinds of heterogeneous instructional settings but not others? What is known about these issues, and how much consensus is there about the cultural values on which schooling must be based?

New Perspectives on School District Research and Evaluation Catherine Lyon Center for the Study of Evaluation University of California, Los Angeles	What is the role of research and development processes in improvement of instruction and learning in basic school subjects in large urban systems? In what form are these processes being applied to the teaching and learning process in urban schools? What is the evidence and how solid are the indications that use of these proceses has resulted in improved student achievement?
Effective Teaching and Learning in Urban Schools Jane A. Stallings and Shirley W. Hentzell Stanford Research Institute	Looking specifically at the interaction of student, teacher, curricular and school organization characteristics, what instructional conditions encourage achievement in basic school subjects? What evidence is there, and how good is it? How are the effects of instructional conditions on student achievement measured? On the basis of existing evidence and trends, what recommendations can be made for improving students' instructional environments?
Mathematics and The Urban Child Marilynn Suydam Center for Science, Mathematics, and Environmental Education The Ohio State University	Mathematics is a school subject which is generally considered basic, insofar as mathematical skills are required for successful adult functioning in our complex society. What evidence is there to indicate which approaches to instruction in mathematics have been successful in increasing achievement? Is there evidence that certain approaches are best suited to the needs of particular populations? Do these effects generalize to other areas of school learning, such as reading, social studies, and science?
Basic Skills in Urban Schools: A View From The Bilingual Classroom Concepcion Valadez Graduate School of Education University of California, Los Angeles	Since the second largest minority population served by many urban school districts is a bilingual population, what is the current state of the evidence about the effects of programs developed especially for bilingual students? What, if anything, does the evidence indicate about the effects of such programs on basic skills achievement?

School-Community Relations and The Urban School Dean Bowles Director Home-School-Community Project Wisconsin Research and Development Center for Cognitive Learning	What do we know or what can we learn about the influence of parental and community involvement in schooling on student achievement? How does parental and community involvement affect the teaching-learning process? If information is not available to answer these questions on even a tentative basis, what types of research are recommended to study this aspect of students' lives more effectively?
Frank Sobol Basic Skills Group National Institute of Education	Are there research and development products which have contributed to improved instruction and increased student achievement in basic school subjects in large urban school systems? What is the evident and how solid are the indications that use of these products has resulted in improved student achivement?
A Summary of The Conference Findings Edmund W. Gordon Director Institute for Urban and Minority Education Teachers College Columbia University	In light of the evidence produced by examination of these issues, what concrete recommendations can be made to school districts concerning the improvement of instruction and learning in basic school subjects? In light of the same evidence, what research and development agenda should be the focus of future study and program development?
Dr. Francis S. Chase, Director of Urban Education Dallas, Texas	Keynote Speaker for Conference Banquet: "Promising Developments in Urban Education"

--Harriet Doss Willis
Director
Urban Education Program

Achievement Tests in Urban Schools: New Numbers

Eva Baker
University of California, Los Angeles

I. INTRODUCTION

It's fun to be popular. In the high school of my past, being popular was an unrealized dream. Hollywood had a surfeit of attractive boys and girls--"youth," one would call them now--and while I never had a chance in that competition, I yearned for acceptance. Instead, I decided good grades and achievement would provide me with my rewards, and after years of good grades and achievements--most of which get counted and weighed rather than reckoned in some larger sense--I find, much to my surprise, that I have chosen a field that is popular. By chance, the popularity has rubbed off on me.

Achievement testing is big news. States mandate it, revered statespersons commend it to us for our spirit, the President likes it, and conferences spring up to discuss and debate it. But as the genie in the bottle reminded us, and as my aging cohort in rock and roll poignantly said, "Sometimes you get what you want, and you don't know what to do with it." I wish I could have made it big without achievement testing.

EVA BAKER is Director of the Center for the Study of Evaluation in the Graduate School of Education, University of California, Los Angeles.

For, while there seems to be a swelling consensus that test taking and giving is an especially legitimate part of school activity, I wonder what it is all for. This paper will explore the legitimacy of the testing of achievement in urban settings. It will consider where this testing comes from, what kinds of tests are administered, how frequently they are given, the purposes to which results are thought to be put, the purposes to which test results are actually put, and some reasons for and against the proposition of testing.

Sources

In public discussion of testing, questions are addressed to determine which kinds of tests are best or how much testing should be done. There appears to be tacit agreement that if we would only expunge tests' sexist, racist, or curricularly irrelevant features, we would be well served by the application of tests. Where does this belief come from? We may, at the outset, plead history. Tests were always used in schools, some say. But a closer review of the history of achievement tests shows them to have been only rarely employed in the late nineteenth century in such fundamental skill areas as spelling. Testing was used as a process to support the authority of the teacher. The teacher already had strong curricular control, and, usually, moral suasion in the classroom. Measurement was used as a formalized adjunct to rather than a substitute for the teacher's judgment. The consequences of testing were relatively intimate and were constrained to the classroom. Momentous decisions, such as promotion to higher grades and to other levels of education, were more often based on economics rather than tests scores, although novels about the effect of posted test scores on English adolescent sensibilities abound.

In contrast, testing is now public, larger than life, and the results bruited about in the media if a school district's performance changes on

some test that obviously deserves everyone's trust. Students are already being threatened or rewarded, depending upon test scores. Where did we gain the faith in testing? Our experience and belief in the power of tests undoubtedly derive from sources other than our familiarity with tests as a part of education.

Our society is competitive, and people have always wanted to know who is better or how this child did as compared to others. Our society is also technological. The most prevalent use of educational technology is based on the research and development (R&D) model of schooling. The early form of this model states that innovation in education should be carefully tested under actual school conditions, that revision and subsequent improvement should occur over time, and that the basis for this feedback and revision is tests. We adopt this mode in development enterprises in aircraft, pharmaceuticals, and even gas-tank configurations. When malfunctions occur in practice, the quality of the testing or the lack of application of its results are impugned. The consumer movement is based on this application of R&D principals, with the expectation that the results of empirical testing will feed back to the developing institution.

Research and development models have been most successful in fields where the validity of the testing instruments could be finely calibrated and clearly described. They have been less effective, in the short run, in the area of social-action programs. An R&D model works well, for instance, in the development of an instructional program planned to account for a limited number of hours of student time, and there is ample evidence to illustrate its success under certain rules. R&D can also be successful when entry to instruction is managed; when conditions of training and program use are constrained; and when the instructor understands how data can control the decision processes involved in the development and use of the program. It is clear that such controls are neither usual, possible, nor desirable in

the creation of large-scale educational action. Yet some of the rhetoric surrounding our use of testing implies that these controls are real. Schools, as they have evolved, just don't fit the test-teaching-test model.

In addition to history, and competition and technology, our belief in tests has derived in part from the acceptance of psychological paraphenalia that prevails in this country. We are given to believe in EST, gamesmanship, parenting, and the reality of a construct embodied as IQ. We believe psychology tests will tell us what we don't already know or can't trust for ourselves. We see this is true from the amount of money spent in psychological activity (self-help or otherwise), in the purchase of periodicals of a popular psychological bent, in the do-it-yourself vocational interest tests, and so on. Accounts of the psychological testing and the widespread testing experience our male, draft-age citizens had have been well documented (Levine, 1976; Buros, 1977; Cronbach & Suppes, 1969). If we can tentatively agree that there are multiple reasons for our investment in testing in schools, we will now consider what tests are used for.

II. EXISTING EVIDENCE

Purposes of Achievement Testing

Most disquisitions on tests begin with some sensible but unalterably pious notion that there is nothing wrong with the idea of testing itself; the problem lies with the application of tests. Paraphasing the Bible: "to every purpose, there is a test." Where we seem to encounter difficulty is in using tests correctly. However, it seems not a little naive to assume that pointing out this penchant for error will correct our problem, when in almost every other sector of modern society, we are plagued with the misapplication of technology and the misuse of good and proper ideas.

In order to assess the extent to which misapplication is a serious issue in testing, a review of the time-honored and more recent uses of tests is appropriate. We will later investigate the types of tests that are proffered by various constituencies to be appropriate to these applications.

In common practice, we test students in schools for the following purposes:

- to select students into appropriate programs within the school;
- to place students diagnostically for particular instruction;
- to provide certification for advancement and/or proficiency of students;
- to permit students to exit early, or on time;
- to select students for programs outside the school--college, for example;
- to evaluate categorical educational programs; to evaluate teaching;
- in given skill areas, to permit comparisons across schools and districts;
- to provide information for the improvement of specific instructional efforts, i.e., formative evaluation;
- to provide mandated information for legislative action;
- to find convenient ways of talking about students;
- to provide feedback to parents and community.

I believe that this list represents a good proportion of the first-order reasons for testing. Second-order reasons--habit, providing activity for the private sector and employement for test specialists--will be excluded, although these uses are probably just as real. The list contains no surprises and should carry a good deal of inherent legitmacy. The purposes as indicated appear both rational and laudable. Such applications allow schools to appear to be performing an important set of functions that should contribute to the credibility of the institution (see Pincus, 1974).

But it is clear that before the kinds of tests that have been employed to serve these various purposes are addressed, the testing intentions should be analyzed somewhat more fully. In reviewing the list, one can see that some purposes have direct effect on students' immediate activities, some have direct or implied effects for later activity, some are provided for administrative reasons, some for reform reasons, and some to provide information to various constituencies. Arraying these purposes against such an analysis provides us with a classification that may be helpful (see below).

FIGURE 1 - TEST PURPOSES BY USE

TYPE 1	DIRECT USES	INDIRECT USES	LEVEL OF APPLICATION
Selection for within school programs	Better service for students	Within school resource allocation	School level
Diagnostic placement	Better service for students	Within school resource allocation	School level
Selection for programs outside of school	Opportunity for students	Service for receiving institution	Outside agency
Certification of competency for advancement or exit (formal)	Personal credentialling	Assessment for quality of service	State of district level
Certification of competency for advancement of exit	Personal credentialling		School level
TYPE 2			
Information for instructional improvement	Change practices	Improve student performance	Classroom, school district, state

Information and feedback for parents	Fulfill expected function	Community support; Student performance improvement	Classroom
Ways to talk about students	Sense of educator expertise	Improvement of instruction; counter negative effects of labelling	School level
TYPE 3			
Evaluation of teaching	Control function	Improvement of service	School level
Comparisons across schools, districts	Trend indication	Public response Improvement of service; Loss in morale	District, State National (NAEP)
Evaluation of categorical programs	Meeting administrative requirements for funding	Legitimation of testing as required activity Assessment of service quality Revision of program	School, District, State, Federal

The classifications presented in Figure 1 represent only rudimentary distinctions about how tests are purportedly used in schools. Testing undoubtedly serves broader gatekeeping functions, reinforces stereotypic views, and, it could be argued, supports inequality. Leaving these issues aside, the uses of test related to specific school operations nonetheless deserve examination.

The type of classifications provided in Figure 1 separate test purposes along lines of basic influence. Type 1 purposes relate to the uses that directly affect students by either guiding the instructional experiences they will receive or by permitting advancement to alternative educational opportunities. Type 2 test purposes serve to affect the activities of

school operation. Tests are used for formative evaluation to improve instruction, for teacher evaluation, and for communication purposes within the school and its local community. Type 3 purposes provide data for broader-based social control, in that the purposes of tests is to provide information about the usefulness of educational programs considered in the broadest perspective. Type 3 purposes for testing are the most usual justifications for test use and, at the same time, probably the most remote in influence on the student.

Test Use in Classrooms

One question that should be raised is whether test information collected for stated purposes is actually used in the predicted ways. Great rhetorical investment has been made in the expectation that testing can improve instruction (see Glaser, 1963). While there is growing evidence that feedback from testing can definitely contribute to the development of instruction and to its effectiveness under controlled conditions (Baker & Alkin, 1973), the claims for the general usefulness of test information to improve what teachers do have yet to be substantially supported (see Block, 1971). Among reasons that this investment in test use is largely unfulfilled may be the nature of tests themselves, a concern to be treated later in this paper. Briefly, the lack of availability of quality instruments sensitive to instructional effects or related clearly to curricula contributes seriously to the problem.

Similarly, one would expect that teachers would need knowledge of testing practices and the ability to make interpretations from data as a prerequisite to their effective use of tests in instructional settings. Some recent work suggests that teachers have neither the knowledge nor predisposition to make use of test information.

Yeh (1978) has conducted work which investigated the knowledge teachers had about testing practices. She found that 50% of her sample were unable to interpret grade equivalent or percentile forms of test reporting, the two most common procedures used to characterize students performance in standarized tests. Goslin (1965) reported on two studies of teaching and testing. His findings suggest that the more training and experience with tests, the greater the test use by teachers--not an altogether suprising conclusion. Yeh found that years of teaching were positively correlated with teachers' reported uses of standarized tests for instructional decision making. The Royal Oak Study (Boyd, Jacobsen, McKenna, Stake, & Yashinsky, 1975) assessed teachers' views on the importance of test giving in their instructional planning. In general, most teachers felt that testing added little to aid them in their instructional endeavors.

Test use for placement and assessment would be expected to be greater when the teacher had on hand a number of instructional alternatives for students. Diagnosis seems sensible only when there are alternative remedies. In Yeh's work, teachers as a group responded weakly (1.8 average where 1 = never, 4 = always) to a query on the uses of tests for instructional decision making in general. No differences were found between teachers who either had or did not have paid aides. (One would expect that paid aides would permit more planning time.) Slight absolute and statistical significant differences were found for teachers in team situations as opposed to those in self-contained classrooms on their reported uses of tests to evaluate the effectiveness of methods and materials. Herman (Note 1) found that teachers' tendency to use either objectives-based or standarized tests for placement purposes was unrelated to the observed variety of types of instructional alternatives in classrooms or to the quality of alternatives available.

The most frequent response of teachers of primary-aged children to the

question, "What do teachers do when test results are poor?" is to drop the skill temporarily, and to provide individual help or move to another approach. Teachers in schools receiving some form of either state or federal compensatory assistance tried another approach more often than teachers in schools without such aid. Approximately 10% of the teachers report dropping the skill and never returning to it. Yeh found that reteaching was the most common teacher response to poor student performance, and that additional practice with the same materials was next preferred, followed by tutoring and additional practice with different materials.

Particular reactions related to teachers in schools with low socioeconomic-status populations present yet a more negative picture. These teachers tend to be less satisfied with the appropriateness of tests; they were less ready to use test results, and tested less frequently, according to Yeh. Goslin (1965) reported that high socioeconomic status is related to testing policies in school; a reasonable inference is that success breeds repetition. Per pupil expenditure is negatively related to the amount of test use, according to Goslin, although positively related to amount of testing--a result that seems a reflection of the general constraints of categorical program evaluation.

Some differences in the types of tests preferred were found in the Baker (1977) study, where teachers in Title One schools preferred curriculum-embedded tests and locally prepared measures to standarized tests. In the Yeh study, teachers in low-SES situations reported they developed their own tests, primarily to permit the more suitable assessment of both student competency and instructional offering.

Tracking the actual use of tests is extremely difficult, and relying exclusively on self-reports of teachers may be unwise. Since teachers are known to group children, perhaps the frequency of regrouping would be an

indicator of test use for individualization. In the Baker study inferences were made about the use of tests for instructional decision making and the provision specifically for individual differences. Teachers were asked how frequently they used and reconstituted groups in reading and mathematics instruction. For these primary-school teachers, approximately 90% used reading groups, 70% used mathematics groups. Mathematics groups were reconstituted more frequently than reading groups, although reported change for both was relatively low. The differential change between reading and mathematics is probably a function of the concreteness with which mathematics proficiency is judged, as opposed to the more diffuse nature of reading ability.

The belief that test-teach-test-reteach procedures can sensibly work in large-scale innovation has been examined in the California Early Childhood Evaluation Program, where teachers are expected to provide data on the individual progress of students to enable individualized instructional planning. Reaction to these procedures has been poor to date, with general compliance by teachers on the provision of data, but annoyance reported, directed to the paperwork requirements mandated by this program. The criticism of documentation decisions related to individualized programs are used to reacting and judging in a fluid, rather linear way, based on formal testing, or simply because a change of habit is required by instructors.

Test Use for Parent Reporting

Teachers in the Yeh and Goslin studies used test scores to assist in parent-reporting requirements particularly as students became older. One must conjecture that the availability of another, presumably objective and valid source of pupil performance helps to diffuse defensiveness between teacher and parent. And while a parent may formulate an opinion of a particular teacher's worthiness, it is hard to impugn the corporate wisdom embedded in

such instruments. Some practical rules for using achievement tests to report student performance in lieu of grades have been described by J. Millman (1970).

Competency-oriented Use of Tests for Certification

A second major grouping of test purposes has developed around the competency-measurement movement. In this orientation, particular, essential compentencies are identified to serve as minimum criteria for advancement or exit from school. Up-to-the minute counts are impossible, but about 33 states had some form of specific standards for promotion or graduation mandated (Pipho, 1977, 1978) as of May, 1978. Calls for a "national test" were also made during the same period. The causes for the interest in competency testing are multiple, but seem to focus on trying to find some way to make the schools explicitly responsible for preparing people to cope with life. Preparing for life seems to be variously construed and may require "survival skills" like checkbook balancing and completing job applications, "basic skills" like reading, writing and arithemetic, career skills, citizenship, and so on. Variations also exist on local control, as in Oregon and California (Hart, 1978), or state control as in Florida (T. Fisher, 1978).

The timing of testing varies, with some large-city schools like those in California having adopted a grade-by-grade testing program for promotion or competency assessment to permit early withdrawal from school or a schedule to allow at least a remedial year. While concern exists about the probity of such testing overall, the effects of such testing on students (see Glass, 1978), the use of input-output models to judge educational efforts (see Wise, 1978), and the arbitrary manner in which competencies are identified and standards set, nonetheless, we have to deal with the reality of these efforts.

In the most benign view, competency testing is important because it tries to assure some equity in the distribution of skills. If everyone can learn, then why shouldn't each person be held responsible for the verification of such competencies? One major inhibition to a calm playing out of this course is the controversy that surrounds the quality of the measures. A second issue is the notion of restricting attention to a very few areas, areas that will reduce the "educational" function of schools to the usually maligned and despised concept of "training." Public view and perspective is being solicited in activities sponsored by the Education Commission of the States and elsewhere (Miller, 1978) but, clearly, technical problems must be solved before consensus can have its desired effect.

In the opinion of many, the compentency-based testing movement is likely to affect students in inner-city regions most strongly. A review of figures on functional literacy and other achievement shows poor performance as a function of socioeconomic status (Breglio, Hinckley, & Beal, 1978; D.L. Fisher, 1978). Furthermore, students with primary languages other than English and those from groups that find themselves in urban centers may be expected to demonstrate competency in English (see Hart, 1978). Since remediation is implied in or specifically tied to most competency-based testing programs, one would expect shifts in resources to follow the curve of poorer performance. In other words, districts with higher incidence of poorer performance should receive more monies. Because the public education budget is unlikely to expand, these resource shifts from surburban schools to the schools of the poor will affect directly that segment of the public supplying the taxes for public school operation. This effect is unlikely to be positive.

Among the problems of attempting to mandate equality of lower-bounded performance rather than equality of opportunity is the difficulty of making the "cut" at which competencies one really wishes to adopt, and which, of

necessity, must be excluded. Recommendations to ease this process have been made (see Brickell, 1978a, 1978b) but the sanguiness with which we can approach such tasks should be moderated by experience. Early attempts at competency-based programs produced expectations that were unrealistic. Teachers, for instance, were unable to distinguish among the importance of over 500 reading objectives (Hoepfner, 1976). Everything was regarded as too important to exclude. To compound the problem, Skager (Note 2) found, for example, that teachers selected as most important to the coming year's academic program those skills in which the students had tested well on the pretest. In addition, to gain consensus, it is also necessary to portray activities at relatively general levels to avoid endless argument and preference, a fact which reduces the understanding and concreteness of the "competency" part of the testing program. Competency testing will be with us for some time; we will need to learn to control it.

Evaluation of Teaching

The idea of using student achievement as an indicator of teacher quality has been in the literature for approximately 20 years (Mitzel & Gross, 1958). It has also been popularized as a point of view that holds everyone can learn and, thus, teacher quality should be assessed partially in terms of the performance of the teacher's pupils. This argument is based on two sets of assumptions: that "all things being equal" is a reasonable proposition; and that teachers are willing to modify their performance in the light of student achievement. Since the "all things being equal" belief in no way corresponds to our understanding of the vast differences in context and operation of schools, the fact that teachers have been demonstrated to improve their teaching based on pupil feedback in college course settings (Popham, 1971) may be only of moderate value. Approaches to the use of student achievement tests to evaluate teachers have varied considerably. Work by Berliner (1976), and by Berliner and Rosenshine (1976) before him,

has adopted the use of "performance tests" by teachers--a lesson or series of lessons with prespecified objectives planned by the teacher. The teacher's ability is then estimated on the basis of the students' performance after these lessons. Among the problems in this area are the evident lack of reliability of these instruments (see Glass, 1973) or the more fundamental question of whether there is consistency of teaching competency by an individual. Findings from work at the Michigan State Institute for Research on Teaching as reported by Shulman (Note 3) suggest that very different skills are needed to teach different subject areas.

In California, an effort to mandate the evaluation of teachers based on student outcome was made during the early 1970s in the form of the Stull Act, which required teachers to identify specific objectives they would use to provide performance data on their teaching. Teachers were permitted to use a wide range of assessment devices, including their own locally prepared measures, commercially available standardized tests, or tests provided with the curriculum. Teachers were also to negotiate their own performance standards with their principal. This was usually translated in terms of the proportion of students who met an arbitrary standard--e.g., 90% of the students will answer correctly 70% of the items. More creative applications developed when some teachers wanted all their students to exceed the class average--a wish often expressed by district-level personnel as well. The Stull Act did not have the desired effect for a variety of reasons, including absence of guidelines and funding, and apparently has died a merciful death. Its only aftermath is the occasional use of the verb "to be stulled," which apparently means to be reviewed using both classroom observation and student-achievement data.

While the desire for teacher evaluation is relatively dormant these days, one can imagine the resurrection of these notions as lower retirement ages and reduced enrollments conspire to make selection and retention issues more

salient. Furthermore, just below the surface of criticism of educational innovations such as Title One, and partly because education is such a personally intensive enterprise, is a deep suspicion directed at the educational endeavor: The idea that teachers may not be doing their job. The competency-based tests that are reported at the student level make possible the reaggregation of data to draw inferences, despite their validity, about the performance of particular teachers and schools. This use of student-achievement tests, if it materializes, will undoubtedly be studied.

Tests in Program Evaluation

Another widely evidenced but controversial use of tests is for the evaluation of educational programs. Tests are used to estimate effects of program innovation under the assumption that if programs were "effective," one first line of evidence would be in the area of increased pupil performance. Often legislation mandates the use of tests as the summary indicator of program effectiveness.

Because considerable disagreement exists as to the appropriateness of tests for assessing program effects, particularly in large-scale federal or state projects, it may be helpful to outline briefly some of the objections:

1. The programs assessed don't have real treatments that may be described; they are funding sources only.
2. The tests used are too general or too specific to capture the educational operations at a particular school.
3. The tests are of low quality.
4. The point at which a test is administered can contribute significantly to whether a program "looks" good or not.
5. Performance outcomes may be manipulated by the way in which students are systematically excluded from tests because of low competency in English.

6. Data analysis procedures obscure importance differences. For instance, using statistical procedures that permit the "equation" of groups using socioeconimic covariates highly correlated with achievement will produce summary results that show statistically induced comparability across groups.
7. The tests are too ad hoc and do not relate to any theory of education.
8. The tests do not provide a continuous picture of progress, but a single time measurement.
9. Nobody uses the data anyway, unless it is to support a political position.
10. Any evaluator can make any set of data look good or bad. Reanalyses of the same data set or use of independent evaluators to complement standard evaluation procedures rarely provide reports with similar conclusions.

Excluding the issue of the nature of the tests themselves, a number of these criticisms can be combined into three clusters: (1) the suspicious-of-data cluster; (2) the suspicious-of-program cluster; and (3) problems-with-administration cluster. Suspicions about the use of data stem from sources other than raw paronoia or affiliation with the new holistic consciousness. In part, they come from the frustration of having something--perhaps a program--measured by instruments that are poorly understood and then analyzed by procedures that are usually arcane. Since evaluation has been used most frequently for comparative, statistical, and measurement purposes, techniques have been employed to heighten distinctions between alternatives, rather than to attempt to represent in any comprehensive way the programs assessed. This suspicion is also fed by the prevalent use of reading (and now mathematics) tests as the exclusive sources of information. The NIE reports that achievement outcomes, particularly in reading and mathematics,

remain the only area in which satisfactory measures are available for formal research (Hill & Rothberg, 1976). Thus, personal beliefs about program effectiveness in comprehensive areas are not verifiable.

Suspicions about programs, particularly about coherent treatment, has plagued large-scale evaluation efforts since the early Title I evaluations (Martin & McClure, 1970). The variations in Title I services retard the ability to characterize a program in even minimally descriptive terms. An alternative strategy has been to attempt to study representative forms of program innovation and perform comparative analyses outcomes related to features of implementation (see Bissell, Note 4; Rivlin & Timpane, 1975; Stallings & Kaskowitz, 1974). In some of these studies descriptions of process variables were related to outcome performance. In a slightly different orientation, Baker (1977) conducted analyses to assess the relationship among process variables and outcomes in the Early Childhood Education Program in California. The intent of this effort was to focus attention on the verification of the hypotheses proferred by the Department of Education and their advisors, hypotheses made operational at one level by program guidelines. Such combination approaches seem plausible in the light of conflicting evidence on which instructional processes are effective under various conditions. Cooley and Lohnes (1976) strongly argued for theory-based evaluation efforts, where process and product are linked so that the findings can contribute to our knowledge in the field of instruction in general.

Administrative problems are frequently linked to complaints about outcome-oriented evaluations; unfortunately, criticisms are all too often deserved. Keesling and Burstein (1977) undertook an audit of testing practices related to the provision of program-evaluation data. Their study identified a series of problems related to the time of test administration (Conklin, Burstein, & Keesling, Note 5; Conklin, 1978), the attrition in

data collected to data actually processed, the processing of data for low-scoring English-speaking students, and the conversions used to equate across standarized test instruments. What should be emphasized is that their object of study was a well-respected but overworked agency, expected to provide evaluation information for a wide variety of constituencies.

The issue of how tests should be given to provide a more fluid picture of progress is debatable. Great cries about the time wasted in testing emerge from educators. On the other hand, evaluation results are perpetually qualified by statements relating to their snap-shot quality, implying that the picture might develop differently were it only taken again. While it would be impossible to progress to a cinematic metaphor by only modestly increasing numbers of testing opportunities, the quality of the inferences would undoubtedly be improved. The benefit of greater confidence would undoubtedly be weighed against the presently perceived nuisance value of the tests.

Other technical problems abound in the interpretation of data gathered under imperfect conditions, including the appropriate analysis unit to include school, class, or student (see Burstein, 1977). While such problems are beyond common knowledge for most educators, their resolution could greatly contribute to the confidence one might place in the testing information collected from students.

Testing for General Community Information

Numerous school districts and state policy bodies mandate tests for the purpose of generally apprising the community of the performance of children. In some instances, statewide assessment programs are used; in others, the standardized achievement test is mandated. The purported function of these tests is to provide overall monitoring of pupil progress

in the administrative area served by policy. School boards want to know how their district is performing, legislatures want to be assured that their support is being reflected in the progress made by the children. Since the news hasn't been all that positive in recent years, this non-specific form of achievement monitoring has taken on a harrassment function. Once a year the education writer for the big city newspaper is sure to be able to point to dismal results. The practice of publishing percentile ranks of schools in Los Angeles has, in the past, augmented copy in real-estate advertisements.

Because the tests are mandated, the public seems to believe that the choice of instrument is rational, and that the quality of school it indicates is a real phenomenon. Changes in tests are always viewed with extraordinary suspicion, especially if students perform better in subsequent years. The interpretation principle embodied in this process is: "Good performance means the test was easy; poor performance means the schools are no good."

Research on the delineation of the organization and function of research, development, and evaluation units in school districts is presently being conducted by Dr. Catherine Lyon. As part of Dr. Lyon's interest, the purposes of school-district level testing will be studied as well as those to whom test results are reported. Information even generally related to these questions is not presently available, except from reports on testing prepared by the General Accounting Office (Problems, 1977). Until descriptive information is gathered about the use to which tests are put in this political context, conclusions about the utility of the effort will, of necessity, have to wait.

Types of Achievement Tests

If we could compare tests sellers to grocery stores--a bizarre concept in

itself--we would be able to characterize the differences in tests that measure achievement. The teacher-made (and sometimes the district-supported) tests might be symbolized by the corner Mom and Pop store. These tests are high on local responsiveness, but often low on sophistication of development techniques. The Educational Testing Service (ETS), with its standarized test product line, might be considered one of the test supermarkets, able to offer a more sophisticated product because it caters to more consumers. Upstarts into this area are the smaller, highly specialized markets--the criterion-referenced tests, if you will. Other variants are the State Assessments, which, it seems, are trying to provide greater economy and less sophistication, while maintaining a level of quality--much like the "plan wrapper" products now popular in supermarkets.

In all these cases, the test maker, like the food supplier, expects the consumer to integrate the product with their experience to come up with satisfying "meals." Yet, tests that are supplied as part of a purchased curriculum often don't allow or require such integration. Curriculum-embedded tests are the fast-food hamburgers of the story--ready to consume, already prepared, but of questionable nutritional value.

Tests obviously come in different forms, packages and from different purveyors. What do their labels mean? A review of the classification of achievement-test types might be helpful at this point. The different kinds of tests have been much discussed in the literature. Sanders and Murray (1976) distinguish among four types of achievement tests: norm-referenced tests; objectives-referenced tests; criterion-referenced tests; and domain-referenced tests. The same four test forms were identified by Harsh (1974).

Norm-referenced tests. Norm-referenced tests are those constructed predominantly to provide a maximum differentiation among students (Novick &

Hambleton, 1973). The purpose of these tests is to rank students' performance with respect to one another on a set of items, thus providing information about distribution. In standarized tests, strong strictures are provided to guide test administration, and raw scores are transformed to allow a standard interpretation,. Common tests are: (1) grade placement, the score achieved by the norm group by children in various school grades; (2) the percentile, a transformation of 1-99, which characterizes the percent of students below a particular level; or (3) the stanine grouping, scores in intervals from 1-9. Standardized tests can be built so that they induce greater spread in student scores above or below the midpoint of the test, depending upon whether finer calibration is desired for higher- or lower-achieving students.

These tests are usually written to provide broad coverage of a content area or skill. Criticisms of norm-referenced tests have been made by a great number of groups because tests have been used to evaluate school programs and the performance of individual students. The basic criticism is that the tests do not tell what a student knows in any descriptive way, but rather place the student (or the school) in a distribution. Another concern is that the "rank" achieved on a test may be related to the similarity or difference between those currently tested and the group the test was "normed on." In general, norm-referenced tests provide little diagnostic information to allow schools or students to improve performance over time. One argument made in favor of these tests is that they provide comprehensive coverage of a particular area, but a paper by Floden (Note 6) showed very little overlap among a number of tests which supposedly measured the same basic area. Thus, a school could look very good or very poor, depending upon how the curriculum actually matched up with the range of skills sampled on the test.

Another potentially telling argument against using such tests for evaluation

is the extent to which the tests are likely to be responsive to the curriculum in use. It has been contended (Popham, 1974) that the need for normed interpretations or comparative information influences the likely sensitivity of the items to program interventions. Arguments in favor of the test-construction process and, therefore, of such tests, have been clearly mounted by Cooley and Lohnes (1976), and Ebel (1972). They support the utility of comprehensive achievement testing that is based on both theoretical and empirical validation procedures. Almost all writers in the field agree that such tests are appropriate for selection of individuals into limited space programs, for instance, where one wishes to provide remediation for the lowest 30% of students. Approximately 90% of schools report using some form of standarized achievement test. (Problems, 1976).

Criterion-referenced tests. Fifteen years ago, the term "criterion-referenced" test came to the fore, propelled by Glaser (1963) and his concern that instruction ought to be measured by procedures different from those provided by the available standardized tests. Since the advent of this article, controversy (and publications) have increased. Criterion-referenced tests are tests that are specifically designed to be influenced by instruction. Lord (1970) writes that our desire under instruction is to have the student able "to answer every test item correctly..." (p. 2).

Criterion-referenced tests are claimed to be appropriate when one wishes to know the extent to which a student has mastered a particular set of information, perhaps for the purpose of providing additional instructional experiments for the student, for advancing the student to a new opportunity, or for correcting and refining the instruction itself. These tests are useful in situations where there are no limits on the number of people who could theoretically succeed on a particular task--"quota-free" in Cronbach and Glaser's terms (1965). Popham and Husek (1969) clarified the difference between development and interpretation techniques for norm- and criterion-referenced testing procedures.

There are still a number of differences in defining what a criterion-referenced test is. Glaser and Nitko (1971) emphasize that "a criterion-referenced test is one that is deliberately constructed to yield measurements that are directly interpretable in terms of specified performance standards" (p. 653). Although Glaser and Nitko expanded their discussion to treat the importance of the need for defining a class or domain of tasks to which these standards are applied, many writers have chosen to focus on the "passing standard" as the definitive characteristic of criterion-referenced tests (Denham, 1975).

The inclination to use passing or cut-off scores as the definitive characteristic of criterion-referenced tests has stimulated interesting research into how such standards are best set to estimate various kinds of error of classification. Work by Harris, Pearlman, and Wilcox (1978); Hambleton and Novick (1973); Shavelson, Block, and Ravitch (1972); and Livingstone (1972) approaches the problem of compound probability models, decision-theoretic approaches, generalizability and reliability frameworks, respectively. Under this interpretation, a passing level or standard is set; there is no clear implication for criterion-referenced test design that is greatly different from the construction of norm-referenced measures. The contribution of criterion-referenced test is their ability to interpret a given student's (or school's) score in terms of "passing" a particular level. Others have interpreted "criterion" to mean "criterion behavior" in the psychological experimentation sense, as pointed out by Glaser and Nitko. With this definition, the emphasis is shifted to the description of what task a learner is able to demonstrate, and "mastery" or passing level is not an imposed requirement. G. Millman (1974) and Popham (1974) have written using this more descriptive interpretation of the term "criterion."

Domain-referenced tests. A logical need would be to find procedures for determining the limits of particular tasks as a precondition to describing

what students can do. Bormuth (1970) discussed the problem of relating achievement tests to particular instruction and recommended the use of test specifications referenced to instruction. Anderson (1972) also dealt with the problem of test construction in reading comprehension, in an effort to provide adequate description of procedures using a set of rules to operationalize the process.

These efforts simply represent a trend toward greater regularity in the test items for specific tasks, an orientation initiated by Hively (1963). Hively developed the idea of domain-referenced testing, with emphasis on the use of a regular and describable procedure in the construction of a set of items, as well as procedures for sampling instances or items from within the domain. The idea seeks to effect the "equivalency" of test items used to assess particular competencies. This equivalency would not only relate to the observable characteristics of the item, but to the psychological process as well. Osburn (1968), and Hively, Patterson, and Page (1968) focus on the creation of these task domains by generating rules or limits to constrain and describe the universe of items to be sampled. These authors attempt to relate their activity to earlier work by Cronbach, Rajaratnam, Nagaswari, and Glaser (1963), and that aspect of generalizability theory that supports random sampling from a defined set of items. Subsequent work by Hively, Maxwell, Rabehl, Sension, and Lundin (1973), Hively (1974), Baker (1974), Popham (1974), and others in this area have attempted to explore procedures for developing such items, procedures for assessing the congruency of items to the domain specified (Baker, Doctorow, & Polin, in press) and forms of interpretation appropriate for classroom and evaluation needs (Hambleton, Swaminathan, Algina, & Coulson, 1978). While the distinction between definitions of criterion-referenced tests might seem academic and arcane, the resolution of this problem has serious implications for how tests are constructed, how students perform on such measures, and--of more importance--how improvement in performance might be approached.

Test bias and domain-referenced tests. One of the serious issues confronting those responsible for test administration is the issue of the biased nature of tests, or the question of whether tests are unfair to particular groups. Most investigations of test bias relate to empirical comparisons of target-group performance on items and subscales of the measure, looking at both the averages of performance attained by other descriptive statistics. Most of the research has focused on bias in selection (Linn, 1973, 1976; Sawyer, Cole, & Cole, 1976; Breglio, Hinckley, & Beal, 1978; Peterson & Novick, 1976), but obviously concern for bias in achievement testing is legitimate as well (Coleman et al., 1966; Comber & Keeves, 1973; Burstein, Keesling, Conklin, & Doscher, 1977). The relationship of achievement tests scores to background factors, such as socio-economic status, fathers' occupation, ranges between .3 and .7, depending upon the sensitivity of the context indicator and whether within- or between-school scores are used. One difficulty with such empirical approaches relates to the problem of disentangling the effects of instruction or lack of it, per se, from the potentially culturally biased aspects of the items themselves.

It is encouraging that writers most known for their scholarship in quantitative areas are moving toward the area of political analysis of the consequences of these decisions (Cronbach, 1976; Novick & Ellis, 1977). In domain approaches, the rules used to generate items are specific: If attention to potential bias is important for the population to be tested, then certain features are either systematically included or excluded in the design of the test. Ratteray (1976) provides a comprehensive review on the issue of cultural bias testing in her work undertaken by the Rand Corporation. At the heart of her suggestions is a procedure for explicitly taking account of alternative cultural contexts in our measurement processes while maintaining concern for achievement level. Although Ratteray does not go beyond the conceptual level, the melding of her analysis and domain-

referenced test procedures seems to contribute to the advancement of the study of bias in ways that extend our present empirical procedures. At present, nonempirical estimates of cultural bias in tests derive from ad hoc content analyses or less rigorous judgments. Adherence to specifications that delimit not only the "content" to be tested, the skill to be demonstrated, but the range of language use appropriate to the respondents and the rules for the selection of a universe of desirable examples, has the potential to transform the problem into manageable terms and will be of special use in analyzing the legal implications of test use.

Objectives-referenced tests. Language was created to capture the new criterion-referenced test movement, and the term "objectives-based," was used during this early period. Objectives-referenced tests represent a less rigorous subset of criterion-referenced tests. In most simple terms, these tests simply meant that test items prepared were referenced to prespecified instructional or behavioral objectives. Objectives-referenced tests are properly placed somewhere in the middle between norm and domain-referenced tests as far as test design goes. Item congruency is judged by assessing the degree of fit of the response format an item has to the statement in objective, instead of the more involved analysis of content parameters as in domain-referenced tests. Most criterion-referenced tests created at either school/district level or by commercial publishers fit the objectives-referenced tests model. The danger in these tests is that they look like domain-referenced tests, they produce scores like domain-referenced tests, but they are created with much less sophistication, and thus, have much less chance to be responsive to instruction with consistency.

Curriculum-embedded tests. Another form of criterion-referenced tests are those that operate as part of purchased curricula. Normally these tests are offered to assist in placement, diagnosis, and to a lesser degree, in final assessment. The tests should be useful to teachers because they are related

to activities supposedly occurring in the classroom, rather than created to assess some more abstract version of a particular subject matter or skill. A review of such measures, however, show them to be susceptible to certain errors:

1. There may not be evidence that these tests have been field tested, a charge that can be leveled at criterion-referenced tests in general.
2. Test items are sometimes prompted; that is, hints are supplied, rather than examples of student performance.
3. Some items are identical to items practiced in instruction.
4. Certain questions are answerable only if one understands some particular definition or specialized language used in th curriculum; idiosyncratic vocabulary may be over-emphasized.
5. Only one or two items may be provided to measure a particular objective.

There are curricula, however, that have integrated some of the precepts of domain-referenced testing in the development of their own materials. For instance, the Southwest Regional Laboratory (SWRL) Communication Skills Program has measures which meet more stringent criteria of domain-referenced testing. (Hansen & Schutz, 1978).

Relationship Among Test Types

At the present time, the distribution of performance when the same children take norm and criterion tests are usually quite similar. These findings may either reduce the importance we attach to the difference in ideology and technique in the tests, or may be taken as evidence that our teaching has yet to become effective, no matter how performance is measured. More

positive results were found in California in the ECE study (Baker, 1977) where domain-referenced tests presented very different pictures of children's performance.

Test Quality

With the burst of energy if not enthusiasm presently devoted to the testing enterprise, and with the recognition of the growing importance placed by policy makers on testing as a way to estimate school effects, one would expect to find some exemplary tests as models. While there are pockets of excellence, the availability of high-quality tests for use in school districts is limited. Reports by the GAO (Program, 1977) document the need for improvement of tests; 89% of the respondents in SEA and LEA positions recommended continued effort in the development of new forms of testing or the improvement of existing measures.

Efforts have been underway to evaluate formally the quality of tests using systematic procedures. The Buros series of Mental Measurement Yearbooks provided a landmark in the evaluation of tests of all sorts. More recently, test evaluation activities have been sponsored by the NIE for norm-referenced tests (Hoepfner, 1976) and for criterion-referenced tests (Walker, 1978). Because any rating process is subject to judgment, the most recent effort by Walker provides an opportunity for clear descriptions of tests and judging categories to include features such as availability of empirical data, practicality, and quality of description of test. Walker reports that most of the criterion-referenced test systems that were judged received low ratings in terms of descriptive power--that, how well one would undersand what the student could do--and in terms of an inadequate number of items to estimate a particular skill area. The consumer aspects of test use have been addressed by Vito Perrone of the National Consortium on Testing in Washington of this year. Clearly, his plea for judgment in the use of tests

was well-spoken. The consumer movement in instruction, one may recall, began in the 1960s with recommendations for technical quality supported by the 1961 Joint Committee on Programmed Instruction and followed by efforts by Komoski of the Education Product Information Exchange (EPIE). The movement has yet to make many inroads in the manner in which instructional materials are selected for use by schools. It may be optimistic to hope that a better fate will be reserved for tests. Buros (1977) called for the availability of technical and developmental data to permit adequate judgment to be made of tests for various uses.

The Test Makers

Who makes tests of these various sorts? There is a wide range of test producers offering their services. Certainly, we are familiar with agencies such as Educational Testing Services and other test publishers. The federal and state government contract to public and private sources for test preparation. Small enterpreneurs, some growing by leaps and bounds, offer their services to create "customized" tests; teacher groups are convened to prepare, with the help of an occasional consultant, tests to assess important competencies. All this activity is quite apart from efforts at state and national assessment, development of instruments for research purposes that often evolve into more general use, or the development of specific measures to meet evaluation requirements.

Walker (1978) found that approximately one-third of the school districts reporting had created their own testing programs using criterion-referenced tests. One might look to more activity of the local development sort, and because of fiscal limitations, less dependence on outside consultants.

Specific Issues in Achievement Tests Related to Urban Students

Issues directly affecting students in urban settings related to achievement

include the various studies undertaken to determine methods of eligibility for compensatory education using either test scores or poverty indices. Studies on this topic are underway, supported by the NIE (1977). A recently completed article on the relationship of big-city school achievement scores and poverty indicators (Meyer & Levine, Note 7) concluded, not surprisingly, that students achieve poorly in neighborhoods with high incidence of poverty and social disorganization. Demographics related to violence, unemployment, and drug use characterize the environment of many urban children (McPartland & McDill, 1977; Levinsohn & Wright, 1976). Harnischfeger and Wiley in their 1975 analysis of the decline of achievement test scores posit that general social conditions seriously contributed to observed decline. They note that in the last 25 years the number of married women with school-aged children has risen from 26 to 51%; that the number of illegitimate births has tripled; and that there has been a substantial (from 10% to 17%) increase of single-parent households. Inadequate replacements for these elements have been made, and the schools apparently have been unable to meet the obligation to educate children who come from gravely disrupted environments. Solutions to problems of testing have been proposed (Heyns, Note 8). Some of these are represented in the line of research related to test-wiseness of inner-city children shown in studies by Diamond, Ayer, Fishman, and Green, 1976; Oakland, 1972; and Millman, Bishop, and Ebel, 1965. It is unlikely that training in a particular skill will turn around the test performance for the urban child.

Frequency and Costs of Testing

How often tests should be given is a problem that can probably be answered in choral response by most teachers: "Less often than they are now." There is a great inefficiency in the number of different tests many children must confront during the school year--particularly children who attend schools that receive various categorical or compensatory-education support. Many

programs mandate specific tests, and children's eligibility makes make them liable for more than one. Then, in addition, a good number of urban school districts have board-mandated testing for district-reporting needs alone. In addition, there may be state assessment, national assessment, and random dissertations to complete. Moreover, under the provisions of P.L. 94-142, handicapped children are to undergo special formal assessment to assure appropriate placement and progress.

What does it cost to test a child? Our estimates on providing a shelf-test battery to a child (amortizing the test over 3 years and including scoring) is about $1.00 per year, per testing occasion. This cost does not include teacher's time, which averages about 50¢ more per child per hour. Nor are the costs of school-district logistics or interpretation time included; nor printing, nor responding to public interviews. And certainly the cost of not teaching something during the testing period has not been computed. If our estimate is relatively low for frequency of testing of an average student in an urban school, and we assume about 2 hours of testing a year, the cost is about $2.00 per student in outlays to test publishers and teachers alone. The $25 million-dollar national figure cited by Anderson and Lesser (1978) seems an underestimate.

When do we give tests? This decision obviously relates to the purposes of the test. Timing of tests seem to relate to whether one is trying to assess growth for a given student evaluation within a year or between year to year growth of cohorts. Testing for instructional purposes probably should take place a few times during the school year, but advocates of more frequent assessment, such as Sension and Rabehl (1974) demonstrate the use of daily testing opportunities. Keesling and Burstein (1977) report differences found in evaluation results as a function of testing dates. Testing for competency certification must be timed relative to the committment one has for remediation. Brickell (1978) recommends schedules that are responsive

to the level of effort possible to marshall for any required remedial effort.

III. SYNTHESIS OF EVIDENCE

The evidence in the area of achievement tests is at once compelling and insignificant. Of most importance is the general agreement emerging in policy on the importance of using achievement indicators as measures of the effects of schooling. The dismaying part of this movement is the paucity of evidence about how tests are actually used.

In brief, research findings indicate the following is true:

1. There is little evidence that decision makers in education--teachers, program managers or policy makers--use the great amount of test data collected annually.
2. Steady increase in testing is being required of students in schools for their own personal certification.
3. Reasons for lack of application are multidimensional but include lack of experience and training with tests and data uses; lack of clarity regarding alternative actions implied by test results; confusion about potential utility and quality of data.
4. Three types of test use were identified: Testing that directly affects the students; testing that is to assist in school operations; testing that is to provide feedback to policy and program directors.
5. Controversy exists over the appropriate test type to use for various purposes.
6. Overall estimates of test quality have not been high.
7. Present emphasis in test development is toward providing descriptive information rather than comparative information for individual-student achievement.

8. The research on test bias has not as yet integrated with newer views of test development.
9. Teachers in low socioeconomic areas are less satisfied with achievement testing procedures than are other teachers.
10. The cost of testing is high.

IV. RECOMMENDATION FOR RESEARCH PROGRAMS

The gaps in data and conclusions create an interlocking series of questions that deserve scholarly attention. While the movement toward testing will undoubtedly not be stalled while programmatic inquiry proceeds, one hopes that the expeditious initiation of such research activities will be supported by all members of the educational community, with the expectation that in the long run, more sensible procedures may be adopted.

A special caution must be added to a call for research to improve our understanding of achievement tests and their proper use in urban settings. The signal is that we should be careful to avoid the expectation that we will be able to translate directly the findings of such research into action plans. Glass (Note 9) reminded us of the dangers of direct extrapolation of research to school practice and his point of view is still cogent. Particularly in the field of achievement, with its character of longitudinal development, our inadequate research procedures (dealing as they usually must with correlational data) limit our inference. Secondly, any findings must be treated carefully, in recognition of the fact that the influences of school achievement are multidimensional in both independent and dependent variables, subject to the effects of time as well as to the unpredictable influences in the environment. Thus, while our interest in unravelling the relationships that affect school achievement in urban settings is a singularly important research program, we must measure our enthusiasm and

constrain the claims we make for instant amelioration based on research.

The research program outlined below is developed from the previous sections. Questions for research will be presented, and occasionally, suggestions for methodology will be provided.

Test Use

Questions:

1. How do teachers or program administrators use tests?
2. Which test types can best improve the instructional process?
3. What kinds of support are required in instruction to assist teachers in the use of tests to modify instructional planning? Are these real rather than fictional matches between tests and curricula?
4. What are the correlates or characteristics of effective test use? Are there urban classroom typologies where such test use has been successful? Can these characteristics be translated so that they are exportable?
5. What are appropriate dependent measures for judging the "effectiveness" of test use?
6. Can research into instructional features be developed so that curriculum-embedded tests--the types most preferred by teachers in urban settings--can be improved, or so that district-imposed tests related specifically to curriculum taught?

Comment

Case-study methodology is obviously best suited to the study of questions related to the actual, rather than reported use of tests. Outcomes of such

research could lead ultimately to the formulation of tentative training requirements for teachers and parents, for the redesign of tests that are included as part of comprehensive curricula identification of critical features in instruction that contribute to important aspects of achievement. Difficulties in this area will lie in the limitation of experimental method for verifying the utility of hypotheses as well as the cost of the research effort itself.

The Achievement Tests

Questions:

1. Can the trade-offs between comparative (norm-referenced) and descriptive (domain-referenced) test strategies be better identified?
2. Does the imposition of specific foci for educational programs, e.g., a limited number of survival skills, seriously moderate or constrain the instructional offering?
3. To what extent does the imposition of standards affect the instructional process?
4. Can technical problems (length of test, timing of measuares, forms of interpretation) be solved in a way that will increase the usefulness of the entire activity?
5. Are relationships betwen norm- and criterion-referenced measures attributable solely to background variables?
6. Can we experiment to determine the contribution of instruction to test performance in natural settings?
7. Can other disciplines be involved to help untangle the inherently confounded ideas of test performance and instructional experience?
8. Can concern for cultural bias of tests be addressed in test design features and subsequently verified empirically? How would one isolate the most critical factors to address?

Costs of Tests

1. What do tests actually cost to administer?
2. Can data collection be reduced so that only one type of measure need be administered and findings, through various sampling and reaggregation processes, be reported to appropriate classroom and administrative units?

Effects of Tests on Students

Questions:

1. What are the effects on students of having passed or not passed competency tests? In the graduation requirement model, what alternatives are available through either the public or private sectors to students who do not pass? What happens to them?
2. How do locally owned models of competency assessment compare to state-administered competency measures? In the long run, is success of programs dependent on the model selected?
3. What is the effect in public support of the school if competency test results indicate high levels of failure?

Beyond Tests

In matters of great moment, it seems appropriate for the nation to pursue multiple paths to reach a desired goal. Certainly, in a pluralistic society, there is great precedent for exploring several probable solutions to pressing problems. One of the major purposes of testing is to assess the effectiveness of schooling. The problems presented by operating schools in many high density, urban regions suggests that estimates of school effectiveness cannot be gleaned from such data. High absence rates, vandalism, attacks of teachers and students, concern for security have transformed the entire set of assumptions that used to guide educational

inferences. Finding Mr. Chips or an exciting learning program cannot provide much impact if the children are not at school. Efforts to mobilize community support for schools are differentially successful. And what about the school that, by chance, serves a population of students who are well-motivated, not plagued by forms of social dislocation, and economically comfortable? Is it fair to assess these schools by the same standard, especially in the second case, where the staff may do very little to contribute to the excellence of achievement test scores?

One alternative that seems appealing is to conduct research and development in the field of educational indicators at the school level. Schools should be responsible for providing the services expected of them for students. They should not be responsible for student achievement when there exist serious, mitigating factors. Discussions of social indicators are usually based on qualtative representations of various products of service effort (see, for instance, Bauer, ed., Social Indicators, 1966). The field of economics has relied on a set of indicators to assess how well the economy is doing. For general, system-level monitoring, an education productivity index should be developed, one largely based on the amalgam and weighting of school services determined to be essential by policy boards. In this way, the idiosyncratic characteristics of location and student population would be excluded from the estimate of whether the school is performing its functions satisfactorily. Instead, the nature of the effort made to make the school a safe but stimulating instructional environment and the competence of the staff and management could be assessed. Factors could be included that either do or do not correlate with typical achievement measures,--e.g. parent involvement--but that are, in themselves, desirable features of school operations. A full research program could be mounted to look at educational productivity, excluding the usual output variables, or at least reducing their contribution. If research on alternative formulation of such an index were conducted with prototypes made available

to school users, the great belief in the power of testing might wane, and expenditures might be sensibly redirected to the school operations and the program evaluation aspects of testing. This redirection would permit educators to focus on the first job of schools: teaching.

V. RECOMMENDATIONS FOR POLICY AND PRACTITIONERS

Since we are enmeshed in an expanding web of testing initiated by the representatives of the people, we should develop some short-term coping strategies to help us do our work.

Test Selection

One important area where practitioners might have clear impact is in the selection of measures. Questions about test quality should be raised, and answers using words other than"reliability" and "validity coefficients" should be expected. To what extent, for instance, does the content on the test match the curriculum in use? If children don't do well on a test, does the test purveyor have recommendations (other than platitudes) for instructional improvement? What is the technical history of the development of the test, especially tests labeled criterion-referenced or domain-referenced tests? What description of the process is provided? Who were the students upon whom the test was tried? For all tests, get copies, not just descriptive information. These questions should be asked, then answered satisfactorily before new tests are secured.

Consulting Firm Use

When consulting firms are employed to assist school districts in the development of tailored tests, determine what credibility the agency has, who has used it before, what process they will use (and if they say they

will write items to match your objectives, get nervous), find out who is actually writing the items (often graduate students with not much training) before contracting. Make sure that if you're paying for the measures, you have exclusive rights to the items.

Education

Begin to do the job of educating yourself and the community on testing and interpretation. Work with the media so that they can begin to understand the purposes of tests, what they do and do not show, and how they are constructed. Coordinated classes of parent and teacher training in understanding the uses and limits of testing might be an appropriate strategy.

Policy

At the policy level, attempt to limit the number of separate testing activities supported in a given unit. Particularly since the use of test data is presently unknown, adopt strategies such as sampling to reduce the burden on teacher and students. Attempt to get permission from project officers in the case of special program funding to use the same measure to provide information for programs with different funding sources. Look into procedures for aggregating data for various reports, rather than administering new and more general tests. Attempt to find ways that the tests used will have maximum benefit for the teachers and students engaged in instruction. When in doubt, test less.

REFERENCE NOTES

Note 1. Herman, J. The relationship of individualized instructional variables and second grade reading, mathematics, and affective outcomes. Dissertation. Los Angeles: University of California, 1978.

Note 2. Skager, R.W. Student entry skills and evaluation of instruction programs: A case study. Paper presented at the annual meeting of the American Research Association, 1969.

Note 3. Shulman, L.S. Paper presented at the CSE Winter Conference on Measurement and Methodology, Center for the Study of Evaluation, University of California, 1978.

Note 4. Bissell, J. The cognitive effects of pre-school for disadvantaged children. Dissertation. Harvard University, 1970.

Note 5. Conklin, J., Burstein, L., & Keesling, J.W. The effects of date of testing and method of interpolation on the use of standardized test scores in the evaluation of large-scale educational programs. Manuscript submitted for publication, 1978.

Note 6. Floden, R.E. Don't they all measure the same thing? Consequences of standardized test selection. Paper presented at the CSE Winter Conference on Measurement and Methodology, Center for the Study of Evaluation, University of California, 1978.

Note 7. Meyer, J.K., & Levine, D.V. Concentrated poverty and reading achievements in five big cities. Paper presented at the annual meeting of the American Educational Research Association, New York, 1977.

Note 8. Heyns, B. Models and measurement for the study of cognitive growth. Working draft prepared for University of Chicago, Educational Finance and Productivity Center (mimeo, n.d.).

Note 9. Glass, G.V. Educational knowledge use. Adapted from an address to the Orientation Program for the U.S. Office of Education, 1969-70. Elkridge, Maryland: Postdoctoral Fellows.

REFERENCES

Anderson, B.D., & Lesser, P. The costs of legislated minimum competency requirements. Phi Delta Kappan, 1978, 59, 606-608.

Anderson, R.C. How to construct achievement tests to assess comprehension. Review of Educational Research, 1972, 42, 145-170.

Baker, E.L. Beyond objectives: Domain-referenced tests for evaluation and instructional improvements. Educational Technology, 1974, 10-21.

Baker, E.L. Evaluation of the California Early Childhood Education Program: Final report. Los Angeles: University of California, Center for the Study of Evaluation, 1977.

Baker, E.L., & Alkin, M.C. Formative evaluation of instructional development. AV Communication Review, 1973, 21.

Baker, E.L., Doctorow, O., & Polin, L. Fuzzy set theory: An alternative approach to descriptive validity at the item level. CSE Technical Report Series. Los Angeles, California: University of California, Center for the Study of Evaluation, in press.

Berliner, D.C. Impediments to the study of teacher effectiveness. Journal of Teacher Education, 1976, 27, 5-13.

Berliner, D.C., & Rosenshine, B.V. Acquisition of knowledge in the classroom. Beginning Teacher Evaluation Study Technical Report Series. San Francisco: Far West Laboratory for Educational Research and Development, 1976.

Block, J.H. Criterion-referenced measurements: Potential. School Review, 1971, 69, 289-298.

Bormuth, J. R. On a theory of achievement test items. Chicago: University of Chicago Press, 1970.

Boyd, J., Jacobsen, K., McKenna, B.H., Stake, R.E., & Yashinsky, J. A study of testing practices in the Royal Oak (Michigan) public schools. Michigan: Royal Oak Michigan School District, 1975.

Breglio, V.J., Hinckley, R.H., & Beal, R.S. Report #2: Students' economic and educational status and selection for compensatory education. Washington, D.C.: Office of Planning, Budgeting and Evaluation, 1978.

Brickell, H.M. Let's talk about minimum competency testing. Denver, Colorado: Educational Commission of the States, 1978. (a)

Brickell, H.M. Seven key notes on minimum competency testing. Phi Delta Kappan, 1978, 59, 589-591. (b)

Buros, O.K. Fifty years in testing: Some reminiscences, criticisms and suggestions. Educational Researcher, 1977, 6, 9-15.

Burstein, L., Keesling, J.W., Conklin, J., & Doscher, M.L. Auditing a large-scale evaluation: The quality of evaluative information for the assessment of program impact and for decision-making. Studies in Educational Evaluation, 1977, 3, 155-168.

Coleman, J.S. et al. Equality of educational opportunity. Washington, D.C.: U.S. Government Printing Office, 1966.

Comber, L.C., & Keeves, J.P. Science education in nineteen countries: an empirical study. New York: John Wiley & Sons, 1973.

Conklin, J. Some methodological issues in using standardized test scores to evaluate large-scale educational programs. CSE Technical Report Series. Los Angeles: University of California, Center for the Study of Evaluation, 1978.

Cooley, W.W., & Lohnes, P.R. Evaluation research in education. New York: John Wiley & Sons, 1976.

Cronbach, L.J. Equity in selection - where psychometrics and political philosophy meet. Journal of Educational Measurement, 1976, 13, 31-41.

Cronbach, L.J., & Gleser, G.C. Psychological tests and personnel decisions. (2nd ed.). Urbana, Illinois: University of Illinois Press, 1965.

Cronbach, L.J., Rajaratnam, Nagaswari, & Gleser, G. Theory of generalizability: a liberalization of realiability theory. British Journal of Statistical Psychology, 1963, 16, 137-163.

Cronbach, L.J., & Suppes, P. (Eds.). Research for tomorrow's schools - disciplined inquiry for education. Report of the Committee on Educational Research of the National Academy of Education. London: MacMillan, Callien MacMillan, 1969.

Denham, C. Criterion-referenced, domain-referenced and norm-referenced measurements: A parallax view. Educational Technology, 1975, 15, 9-13.

Diamond, J.J., Ayer, J., Fishman, R., & Green, P. Are inner city children test-wise? Journal of Educational Measurement, 1976, 14, 39-45.

Ebel, R.L. Criterion-referenced measurements: Limitations. School Review, 1971, 69, 282-288.

Fisher, D.L. Functional literacy and the schools. Washington, D.C.: National Institute of Education, 1978.

Fisher, T.H. Florida's approach to competency testing. Phi Delta Kappan, 1978, 59, 599-601.

Glaser, R. Instructional technology and the measurement of learning outcomes. American Psychologist, 1963, 18, 519-521.

Glaser, R., & Nitko, A.J. Measurement in learning and instruction. In R.L. Thorndike (Ed.), Educational Measurement, Washington: American Council on Education, 1971, 625-670.

Glass, G.V. Statistical and measurement problems in implementing the Stull Act. In N.L. Gage (Ed.), Mandated Evaluation of Editors: a Conference on California's Stull Act. Stanford, California: Stanford Center for Research and Development in Teaching, 1973, 53-98.

Glass, G.V. Minimum competence and incompetence in Florida. Phi Delta Kappan, 1978, 59, 602-604.

Goslin, D.A. The use of standardized tests in American secondary schools and their impact on students, teachers and administrators. New York: Russell Sage Foundation, 1965.

Hambleton, R., & Novick, M. Toward an integration of theory and method for criterion-referenced tests. Journal of Educational Measurement, 1973, 10, 159-170.

Hambleton, R.K., Swaminathan, H., Algina, J., & Coulson, D.B. Criterion-referenced testing and measurement: A review of technical issues and developments. Review of Educational Research, 1978, 48, 1-47.

Hanson, R.A. & Schutz, R.E. A new look at schooling effects from programmatic research and development. In D. Mann (Ed.), Making change happen? New York: Columbia University Press, 1978.

Harnischfeger, A., & Wiley, D.E. Achievement test score decline: Do we need to worry? St. Louis: CEMREL, Inc., 1975.

Hart, G.K. The California pupil proficiency law as viewed by its author. Phi Delta Kappan, 1978, 59, 592-595.

Hill, P. & Rothberg, I. Evaluating compensatory education: An interim report on the N.I.E. Compensatory Education Study. Washington, D.C.: National Institute of Education, 1976.

Hively, W. Defining criterion behavior for programmed instruction in elementary mathematics. Cambridge: Harvard University, 1963.

Hively, W. Introduction to domain-referenced testing. Educational Technology, 1974, 6, 5-10.

Hively, W., Maxwell, G., Robehl, G., Sension, D., & Lundin, S. Domain referenced curriculum evaluation: A technical handbook and case study from the Minnemast Project. CSE Monograph Series in Evaluation, No. 2. Los Angeles: University of California, 1973.

Hively, W., Patterson, H., & Page, S. A universe of defined system of arithmetic achievement test. Journal of Educational Measurement, 1968, 5, 275-290.

Hoepfner, R. (Ed.). CSE elementary school test evaluations (Rev. ed.). Los Angeles: University of California, Center for the Study of Evaluation, 1976.

Keesling, W., & Burstein, L. An audit report of the activities of the California State Department of Education related to the Early Childhood Education Program. Early Childhood Education Evaluation (Vol. 2). Los Angeles: University of California, 1977.

Levine, M. The academic achievement test: Its historical context and social functions. American Psychologist, 1976, 228-238.

Levinsohn, F.H., & Wright, D.B. (Eds.). School desegration: Shadow and substance. Chicago: The University of Chicago Press, 1976.

Linn, R.L. Fair test use in selection. Review of Educational Research, 1973, 43, 139-162.

Linn, R.L. In search of fair selection procedures. Journal of Educational Measurement, 1976, 13, 53-58.

Lord, F.M. Some test theory for tailored testing. In W. Holtzman (Ed.), Computer-assisted instruction, testing and guidance. New York: Harper and Row, 1970.

Martin, R., & McClure, P. Title I of ESEA: Is it helping poor children? Washington: Washington Research Project of the Southern Center for Studies in Public Policy & the NAACP Legal Defense and Educational Fund, 1970.

McPartland, J.M., & McDill, E.L. Violence in schools. Lexington, Massachusetts: Lexington Books, 1977.

Miller, B.S. (Ed.). Minimum competency testing: A report of four regional conferences. St. Louis: CEMREL, Inc., 1978.

Millman, G. Criterion-referenced measurement. In W.J. Popham (Ed.), Evaluation in education: Current applications. Berkeley, California: McCutchan Publishing, 1974.

Millman, J. Reporting student progress: A case for a criterion-referenced marking system. Phi Delta Kappan, 1970, 52, 226-230.

Millman, J., Bishop, C.H., & Ebel, R. An analysis of test-wiseness. Educational and Psychological Measurement, 1965, 25, 707-726.

Mitzel, H.E., & Gross, C.F. The development of pupil growth criteria in studies of teacher effectiveness. Educational Research Bulletin, 1958, 37, 178-187, 205-275.

Novick, M.R., & Hambleton, R.K. Toward an integration of theory and method for criterion-referenced tests. Journal of Educational Measurement, 10, (3) Fall 1973, 159-170.

Oakland, T. The effects of test-witness materials on standardized test performance of preschool disadvantaged children. Journal of School Psychology, 1972, 10, 355-360.

Osburn, H.G. Item sampling for achievement testing. Educational and Psychological Measurement, 1968, 28, 95-104.

Petersen, N.S., & Novick, M.R. An evaluation of some models for culture-fair selection. Journal of Educational Measurement, 1976, 13, 3-30.

Pincus, J. Incentives for innovation in public schools. Review of Educational Research, 1974, 4, 113-144.

Pipho, C. Update VII: Minimal competency testing. Denver, Colorado: Education Commission of the States, 1977.

Pipho, C. Minimum competency testing in 1978: A look at state standards. Phi Delta Kappan, 1978, 59, 585-588.

Popham, W.J. Performance tests of teaching proficiency: Rationale, development and validation. American Educational Research Journal, 1971, 8, 105-17.

Popham, W.J. Selecting objectives and generateing test items for objective based tests. CSE Monograph Series in Evaluation. Los Angeles: University of California, Center for the Study of Evaluation, 1974, 13-26.

Popham, W.J., and Husek, T.R. Implications of criterion-referenced measurement. Journal of Educational Measurement, 1969, 6, 1-9.

Problems and needed improvements in evaluating Office of Education Programs. Report to the Congress by the Controller General of the United States. Washington, D.C.: GAO, HRD 76-165, September 1977.

Ratteray, J.D. The testing of cultural groups: A paradigmatic analysis of the literature on testing and a preposition. Santa Monica, California: The Rand Corporation, 1976. (ED 113-371).

Rivlin, A.M., & Timpane, P.M. (Eds.). Planned variation in education: Should we give up or try harder? Washington, D.C.: The Brookings Institute, 1975.

Sanders, J.R., & Murray, S. Alternatives for achievement testing. Educational Technology, 1976, 17-23.

Sawyer, R.L., Cole, N.S., & Cole, J.W.L. Utilities and the issue of fairness in a decision theoretic model for selection. Journal of Educational Measurement, 1976, 13, 59-76.

Sension, D.B., & Rabehl, G.J. Test-item domains and instructional accountability. In W. Hively (Ed.), Domain referenced tests. Educational Technology Press, 1974.

Shavelson, R.J., Block, J.H., & Ravitch, M.M. Criterion-referenced testing: Comments on reliability. Journal of Educational Measurement, 1972, 9, 133-158.

Skager, R.W. The great criterion-referenced test myth. CSE Technical Report No. 95. Los Angeles: University of California, Center for the Study of Evaluation, 1978.

Stallings, J.A., & Kaskowitz, D.H. Follow-through classroom observation evaluation, 1972-1973 (SRI Project URU-7370). Menlo Park, California: Stanford Research Institute, 1974.

Walker, C.B. Criterion-referenced tests evauation book. Los Angeles: University of California, Center for the Study of Evaluation, 1978, in press.

Wise, A.E. Minimum competency testing: Another case of hyper-rationalization. Phi Delta Kappan, 1978, 59, 596-598.

Yeh, J.P. Test use in schools. CSE Technical Report Series. Los Angeles: University of California, Center for the Study of Evaluation, 1978, in press.

OTHER TITLES IN THE SERIES

Volume 1
Compensatory Education: A Congressional Perspective
Christopher T. Cross, the U.S. House of Representatives

Volume 2
Effective Teacher-Support Systems to Improve Instructional Conditions in Schools
Joseph D. Liggins, Houston Independent School District

Volume 3
Assessment That Respects Complexity in Individuals and Programs
Judah L. Schwartz, Massachusetts Institute of Technology

Volume 4
Achievement Tests in Urban Schools
Eva Baker, University of California, Los Angeles

Volume 5
Instructional Ingredients for the Development of Beginning Reading Competence
Isabel L. Beck, University of Pittsburgh

Volume 6
A Discussion of the Literature and Issues Related to Effective Schooling
Ronald Edmonds, Harvard University and New York City Public Schools

Volume 7
Summer Drop-Off and the Effectiveness of Compensatory Instruction
Paul T. Hill, The Rand Corporation

Volume 8
The Educational Promise of Cultural Pluralism
Vera John Steiner and Larry Smith, The University of New Mexico

Volume 9
New Perspectives on School District Research and Evaluation
Catherine Lyon, University of California, Los Angeles

Volume 10
Effective Teaching and Learning in Urban Schools
Jane A. Stallings and Shirley W. Hentzell, SRI International

Volume 11
Mathematics and the Urban Child: A Review of Research
Marilyn N. Suydam, The Ohio State University

Volume 12
Basic Skills in Urban Schools: A View from the Bilingual Classroom
Concepcion M. Valadez, University of California, Los Angeles

Volume 13
School-Community Relations and Student Achievement
John E. Ingram, Jr. and Lois I. Bartels, St. Louis Public Schools and CEMREL, Inc.

Volume 14
What Do We Know About Teaching and Learning In Urban Schools: A Summary of the Conference Findings
Edmund W. Gordon, Columbia University